CHALMERS
UNIVERSITY OF TECHNOLOGY

Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data

Citation for the original published paper (version of record):
Sokolova, E., Ivarsson, O., Lillieström, A. et al (2022). Data-driven models for predicting microbial water quality in the drinking water source using E.
coli monitoring and hydrometeorological data. Science of the Total Environment, 802.
http://dx.doi.org/10.1016/j.scitotenv.2021.149798

Science of the Total Environment 802 (2022) 149798

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Data-driven models for predicting microbial water quality in the drinking water source using *E. coli* monitoring and hydrometeorological data

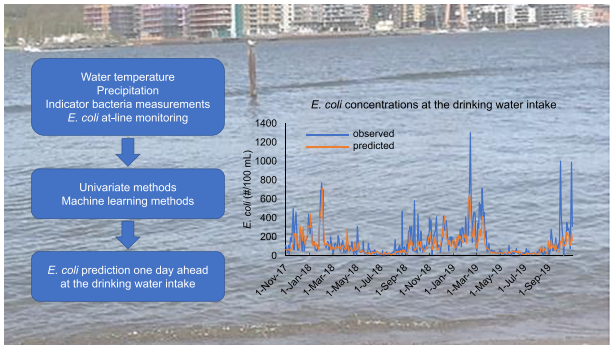
Ekaterina Sokolova [a,*], Oscar Ivarsson [b], Ann Lilliestråle [b], Nora K. Speicher [b], Henrik Rydberg [c], Mia Bondelind [a]

[a] Chalmers University of Technology, Department of Architecture and Civil Engineering, Sweden
[b] Chalmers University of Technology, Department of Computer Science and Engineering, Sweden
[c] City of Gothenburg, Department of Sustainable Water and Waste, Sweden

HIGHLIGHTS

- For predicting *E. coli*, data-driven models of different complexity were evaluated.
- Models with multiple predictors outperformed univariate models in predicting *E. coli*.
- Important predictors were temperature, microbial concentrations, and precipitation.
- Models help interpret what concentrations are expected and identify unexplained peaks.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 April 2021
Received in revised form 8 July 2021
Accepted 16 August 2021
Available online 21 August 2021

Editor: Yucheng Feng

Keywords:
Artificial intelligence
Drinking water
E. coli
Machine learning
Microbial water quality

ABSTRACT

Rapid changes in microbial water quality in surface waters pose challenges for production of safe drinking water. If not treated to an acceptable level, microbial pathogens present in the drinking water can result in severe consequences for public health. The aim of this paper was to evaluate the suitability of data-driven models of different complexity for predicting the concentrations of *E. coli* in the river Göta älv at the water intake of the drinking water treatment plant in Gothenburg, Sweden. The objectives were to (i) assess how the complexity of the model affects the model performance; and (ii) identify relevant factors and assess their effect as predictors of *E. coli* levels. To forecast *E. coli* levels one day ahead, the data on laboratory measurements of *E. coli* and total coliforms, Colifast measurements of *E. coli*, water temperature, turbidity, precipitation, and water flow were used. The baseline approaches included Exponential Smoothing and ARIMA (Autoregressive Integrated Moving Average), which are commonly used univariate methods, and a naive baseline that used the previous observed value as its next prediction. Also, models common in the machine learning domain were included: LASSO (Least Absolute Shrinkage and Selection Operator) Regression and Random Forest, and a tool for optimising machine learning pipelines – TPOT (Tree-based Pipeline Optimization Tool). Also, a multivariate autoregressive model VAR (Vector Autoregression) was included. The models that included multiple predictors performed better than univariate models. Random Forest and TPOT resulted in higher performance but showed a tendency of overfitting. Water temperature, microbial concentrations upstream and at the water intake, and precipitation upstream were shown to be important predictors. Data-driven modelling enables water producers to interpret the measurements in the context of what concentrations can be expected based on the recent historic data, and thus identify unexplained deviations warranting further investigation of their origin.

* Corresponding author at: Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.
E-mail address: ekaterina.sokolova@chalmers.se (E. Sokolova).

https://doi.org/10.1016/j.scitotenv.2021.149798
0048-9697/

1. Introduction

Rapid changes in microbial water quality in surface waters complicate the optimisation of the water treatment at the drinking water treatment plants. If not treated to an acceptable level, microbial pathogens still present in the drinking water can result in severe consequences for public health as they may cause waterborne disease outbreaks (WHO, 2017). Waterborne outbreaks could also cause lower trust from consumers, increase their perceived risk, and decrease their acceptance for drinking water (Bratanova et al., 2013). Changes in the microbial water quality of surface water are often caused by heavy rainfall leading to wastewater discharges from sewer systems and increased runoff from grazing areas and agricultural fields. Laboratory analyses of microbial water quality are available only with a delay in time. It is therefore of value to predict and forecast the microbial concentrations in the incoming water to the drinking water treatment plant to be able to implement measures, e.g., use an alternative water source or optimize the treatment processes.

Microbial water quality can be predicted using data-driven models. For example, artificial neural networks have been successfully applied to predict microbial water quality in terms of compliance with recreational water quality regulations (Avila et al., 2018; Choi and Bae, 2018; Laureano-Rosario et al., 2019; Vijayashanthar et al., 2018), often alongside various regression methods (Mas and Ahlfeld, 2007; Motamarri and Boccelli, 2012; Thoe et al., 2012, 2014, 2015), as well as classification trees (Avila et al., 2018; Stidson et al., 2012). In the context of drinking water supply, there are also some recent attempts to predict the concentrations of microorganisms using, e.g., zero-inflated regression models, random forest regression model, adaptive neuro-fuzzy inference system, and Gaussian process for machine learning (Mohammed et al., 2017a, 2017b, 2017c, 2018), as well as of other pollutants (Asheri Arnon et al., 2019; Samanipour et al., 2019; Speight et al., 2019; Stevenson and Bravo, 2019). A recent review (Francy et al., 2020) provides several examples when data-driven methods are used in operational nowcasting systems for public notification and water management. In the literature cited above, data-driven models are based on monitoring data for target variables, most often faecal indicators, but also on other explanatory variables, for which the datasets are easier to obtain. The explanatory variables used in these data-driven approaches can be divided into meteorological (most often precipitation, but also solar radiation and wind conditions), hydrological (most often river flow, but also sea level or tide conditions), and water quality (most often turbidity, but also water temperature, pH, salinity, electrical conductivity, colour, alkalinity, and past microbial concentrations).

Water producers need to manage their operation in response to rapid variations in microbial water quality. Currently, this is largely done based on monitoring data on faecal indicator concentrations. However, monitoring cannot fully capture the high variability of microbial concentrations over time and in space, and the laboratory results for microbial concentrations are often available only after the water has already been treated, due to the time required for microbial analyses. In case of the river Göta älv, the drinking water source for the city of Gothenburg, frequent microbial data (two measurements per day) are obtained using at-line monitoring. Moreover, earlier studies have shown that precipitation drives variations in microbial water quality in this river (Åström et al., 2007; Tornevi et al., 2014). Here, we explore whether at-line monitoring of *E. coli* can be combined with hydrometeorological data in a data-driven model to predict concentrations of *E. coli* at a drinking water intake. The novelty of this paper lies in using the uniquely extensive dataset of frequent at-line monitoring of *E. coli* to develop a data-driven model.

The aim of this paper was to evaluate the suitability of data-driven models of varying complexity for predicting the concentrations of *E. coli* at the drinking water intake for the city of Gothenburg in Sweden. The objectives were to (i) assess how the complexity of the model affects the model performance; and (ii) identify relevant factors and assess their effect as predictors of *E. coli* levels.

2. Methodology

2.1. Study area and data

Göta älv is a river that drains Lake Vänern into the strait Kattegat at the city of Gothenburg on the west coast of Sweden (Fig. S1). The total catchment area of the river Göta älv is 50,233 km^2, which constitutes approximately 10% of the area of Sweden. The part of the catchment area that is located downstream of Lake Vänern is approximately 3500 km^2. The length of the river between the outflow from Lake Vänern and the mouth of the river is 93 km. The vertical drop of the river is approximately 44 m. The water flow in the river Göta älv is regulated by several hydropower stations (Fig. S1) and varies strongly; the mean and the maximum water flows are 550 and 1000 m^3/s, respectively. The transport time between the outflow from Lake Vänern and the mouth of the river is between 1.5 and 5 days.

The river is used as a water source for the drinking water supply of 700,000 consumers in several municipalities, including Gothenburg with 500,000 consumers. Between Lake Vänern and the water intake for the city of Gothenburg (Fig. S1) the river receives wastewater from approximately 100,000 persons. Approximately 95% of this wastewater is treated at municipal wastewater treatment plants (WWTPs), while 5% is treated by on-site sewer systems.

The dataset used for model development (Table 1) consisted of laboratory measurements of *E. coli* (SS-EN ISO 9308-2:2014) and total coliforms (SS-EN ISO 9308-2:2014), Colifast measurements of *E. coli* (Colifast AS fluorometric monitoring of β-glucuronidase activity), water temperature, turbidity (SS 028125), precipitation (provided by the Swedish Meteorological and Hydrological Institute), and water flow (provided by the power company Vattenfall). The water quality data were provided by the city of Gothenburg.

For model development, the time period 3 Apr 2012–30 Oct 2019 was selected. The data were split into a training dataset (3 Apr 2012–30 Oct 2017) and a test dataset (1 Nov 2017–30 Oct 2019). For the target variable *E. coli* laboratory measurements at Lärjeholm, the dataset included 872 and 341 observations for the training and test periods respectively.

Plotting the target variable revealed annual seasonal behaviour (Fig. S3) with extreme values and large variation during the winter months. The seasonal pattern is complex due to irregular sampling frequency as well as varying timing, duration, and magnitude of increased concentrations. It was therefore assumed that the seasonal variation could be explained using external predictors showing a similar seasonality, e.g., water temperature. While precipitation and water flow data were available at regular intervals (daily or hourly values), water quality data were available at irregular intervals with time between observations up to several days. To combine regular and irregular time-series, the irregular time-series were sampled at a daily basis leaving days with no observations empty, and lags of observations were used as input to the models. The exception was the VAR model that requires the input data to be regularly spaced, thus, for this model, forward filling was performed for the irregular features. Data imputation for the Colifast measurements was performed by using the values from the other location if available, otherwise forward fill was performed. As the Colifast measurements were provided as ordinal categorial values (<50, 50, 100, 200, 400, >400), to be used in the model, they were transformed to numerical values keeping the same order.

2.2. Model set-up

The focus in this project was on forecasting *E. coli* levels one day ahead. Two approaches were used, with original values for numerical

Table 1
Summary of the dataset used for model development.

Type	Location[a]	Time resolution	Unit	5th perc.	50th perc.	95th perc.	% days missing
E. coli Lab	LAE	Avg. 2 d 6 h	#/100 mL	10	86	483	56
	GA	Avg. 3 d 3.5 h	#/100 mL	10	90	567	68
E. coli Colifast	LAE	Twice a day	#/100 mL	NA[b]	NA	NA	4
	GA	Twice a day	#/100 mL	NA	NA	NA	9
Total Coliforms	LAE	Avg. 2 d 4.5 h	#/100 mL	120	445	3325	54
	GA	Avg. 3 d 2.5 h	#/100 mL	110	470	3900	68
Precipitation[c]	GBG	Daily sum	mm	0	0	13.0	0
	VB	Daily sum	mm	0	0	11.8	0
	KR	Daily sum	mm	0	0	15.1	0
Water flow	LE	Daily mean	m^3/s	206	550	900	0
	GBG	Daily mean	m^3/s	132	152	225	0
Water temp.[c]	LAE	Avg. 2 d 1 h	°C	0.8	8.8	18.7	52
Turbidity	LAE	Avg. 2 d 7.5 h	FNU	3.3	6.25	16.4	58

[a] LAE – Lärjeholm, GA – Garn, GBG – Gothenburg VB – Vänersborg, KR – Komperöd, LE – Lilla Edet. The locations are shown in Fig. S1.
[b] NA – not available, as the Colifast measurements were provided as ordinal categorial values (<50, 50, 100, 200, 400, >400 #/100 mL).
[c] The data for precipitation at GBG and water temperature at LAE are visualised in Fig. S2.

microbial features (*E. coli* Lab and Coliforms) and with Log-scaling using $Log_{10}(1 + x)$ transformation. When the original values were used, hyperparameter optimisation for some of the models included logarithmic transformations on the target variable, but the values were transformed back to their original scale before evaluation.

To compare methods of different complexity, both univariate and multivariate methods were applied. The univariate approaches included Exponential Smoothing and ARIMA (Autoregressive Integrated Moving Average), which are common methods for modelling and forecasting time-series data; also, a naive baseline that used the previous observed value as its next prediction was included. The multivariate approaches included methods from the machine learning domain: LASSO (Least Absolute Shrinkage and Selection Operator) Regression and Random Forest, as well as a tool for optimising machine learning pipelines – TPOT (Tree-based Pipeline Optimisation Tool). These methods represent different levels of complexity: LASSO Regression is an extension of Linear Regression using regularisation, Random Forest can model non-linear relationships between input features and the target variable, and TPOT can produce pipelines with multiple models stacked on top of each other. To complement the machine learning approaches and assess if similar performance can be achieved, a multivariate autoregressive model VAR (Vector Autoregression) was included.

To make these multivariate approaches applicable for a time-series problem, feature engineering with lagged features as input was used. Lagged features are features containing data from prior time steps. The lags were defined based on observations, which for some features were regular (daily) and for some features irregular. The maximum lag that was used for each feature was selected based on domain knowledge and previous publications (Tornevi et al., 2014). The predictors were: water temperature with a lag of 1 observation; laboratory measurements of *E. coli*, total coliforms and turbidity with lags of 1–3 observations; water flow with lags of 1-5 days; precipitation with lags of 1-6 days; *E. coli* Colifast with lags of 1–10 observations.

The performance of the different models was measured using common error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), and Symmetrical Mean Absolute Percentage Error (SMAPE) – an accuracy measure based on percentage errors allowing for observation values that are zero. The performance was measured on the training and validation splits during the cross validation, as well as the unseen test data. The model performance was also inspected visually by plotting forecasted predictions along with the true observations.

To interpret the contribution of different features to the target variable, LASSO and Random Forest methods were used. LASSO produces coefficients of the standardised variables that show the effect each variable has on the target variable. The LASSO regularisation path can be used to show what effect different values of the regularisation parameter alpha have on the coefficients (Fig. S4). Random Forest can be used to extract feature importance based on how much each feature contributes to decreasing the impurity in the trees; this can be used to rank the different features. However, this ranking is performed using the training data, and may be less representative for the unseen data in case the model is overfitting.

The applied models are described below, and the parameterisation is summarised in Table S1.

2.2.1. Naive

The naive model uses the most recent observed value as its next prediction. This simple approach was used as a baseline for comparing the performance of more complex models.

2.2.2. Exponential Smoothing

Exponential Smoothing was selected as a simple extension of the Naive baseline. This method considers past observations and weighs their importance over time in an exponentially decreasing manner, so that the most recent observation has the highest weight. In this study, a variant called Simple Exponential Smoothing was used. This approach is mostly suitable for data without a clear trend or seasonality. This method was evaluated despite the seasonality in the microbial data to get an indication of its overall usefulness. A promising result could motivate an extension of the method to better capture seasonality. The method has a single parameter α that decides the smoothing factor, i.e., how much importance the model should allocate to its most recent value. The parameter α was optimized automatically using the implementation of Exponential Smoothing in the Python package statsmodel (https://www.statsmodels.org/) (Seabold and Perktold, 2010).

2.2.3. Autoregressive Integrated Moving Average (ARIMA)

ARIMA (Box and Jenkins, 1976) was selected since it is one of the most common methods for forecasting time-series. ARIMA is a type of linear regression model that is auto-regressive (the "AR" term), meaning that it uses previous values of the target variable to make its predictions. The term "Integrated – I" denotes the use of differencing to make the mean and variance consistent over time, i.e., to make the time series stationary, resulting in a more robust model. The term "Moving Average – MA" means that the model forecasts a value using the model's past errors, i.e. attempts to correct the deviations between the past predicted and true values. The parameters (p, d, q) are used for selecting the order of the AR, I and MA terms in the model. The p and q parameters are optimized with the Akaike Information Criterion (AIC) (Akaike, 2011), and the parameter d is selected based on the KPSS-test (Kwiatkowski et al., 1992) for stationarity. This was performed using the implementation of ARIMA in the Python package pyramid (https://alkaline-ml.com/pmdarima) (Smith, 2017).

2.2.4. Vector Autoregression (VAR)

VAR was selected because it is a generalisation of the univariate autoregressive model that allows for multivariate time-series. Each variable is modelled through a linear equation including its lagged values, the lagged values of the other variables, and an error term. The lag parameter p is equal for each input variable and is selected using AIC. Since input data need to be regularly spaced, imputation by forward filling was used in order to obtain daily values for the irregular features. VAR was performed using the Python package statsmodel (https://www.statsmodels.org/) (Seabold and Perktold, 2010).

2.2.5. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO regression is a method for identifying linear relationships in the data while avoiding overfitting through regularisation. The result is a mathematical equation that defines the target variable as a function of the predictor variables. LASSO is also robust against multicollinearity in the data and facilitates best feature selection. In general, regularisation is implemented by adding a penalty to the best fit of the training data, in order to make the model generalise better on unseen test data. Regularisation also restricts the influence of predictor variables by disallowing their coefficients to grow too large. LASSO uses the L1 regularisation technique, which allows coefficients to become zero. In this way, LASSO can be used to automatically select relevant features. LASSO uses a parameter α that defines how much regularisation should be applied, thus, an α value of 0 would mean performing ordinary linear regression. The parameter α was optimized using a 5-fold cross validation on the training data with an expanding window setup not to violate the temporal dependency in the data. To find the best parameters, a grid search was performed, and the best model was found by averaging the mean absolute error across the folds. Other hyperparameters that were optimized were whether to perform Log_{10}-transformation on the target variable and whether to perform power scaling on the predictors. LASSO was performed using the Python package scikit-learn (https://scikit-learn.org/) (Pedregosa et al., 2011).

2.2.6. Random Forest

Random Forest was selected in order to evaluate whether better performance can be achieved with a more complex model, as it has been showed to work for a great number of tasks on various datasets. Also, this method can produce a feature importance score to see how much each feature contributes to the predicted values. Random forest is a popular ensemble learning method, which works by combining multiple decision trees and outputting their average prediction. A decision tree model builds a tree like structure from the observed data, where observations on the features of the data are represented by the branches of the tree, and the conclusions are represented in the leaves. Compared to individual decision trees, the Random Forest method counteracts overfitting of the data and errors due to bias. In this study, the same cross validation setting as for LASSO was used, and the following hyperparameters specific to Random Forest were tuned: maximum depth, minimum sample per split, minimum sample per leaf, maximum ratio of features, and with or without bootstrapping. Random Forest was performed using the Python package scikit-learn (https://scikit-learn.org/) (Pedregosa et al., 2011).

2.2.7. Tree-based Pipeline Optimisation Tool (TPOT)

TPOT was selected because if offers an estimation of the performance that is achievable using much more complex model architectures. It is an automated machine learning library, which uses genetic programming to optimize machine learning pipelines. TPOT automates feature selection, model selection and parameter optimization in order to find the best predictive model of the data at hand. The resulting model often tends to become quite complex and hard to interpret. TPOT was used through the Python package TPOT (https://epistasislab.github.io/tpot/) (Le et al., 2020).

3. Results and discussion

3.1. Model performance

All applied models performed better than the Naive baseline (Table 2), indicating that there are patterns in the data that can be used for a prediction for the next day. Exponential Smoothing and ARIMA showed similar performance; MAE and SMAPE showed better performance of Exponential Smoothing, RMSE and R^2 showed better performance of ARIMA. The fact that MAE and SMAPE are more sensitive to small errors and RMSE is more sensitive to large errors, indicates that ARIMA performs worse during periods with lower *E. coli* levels and could potentially be improved by a seasonal component in the model (Fig. 1).

External predictors improved model performance, with the best performance for the original data achieved by the TPOT library (Table 2, Fig. 1). However, the TPOT method produces complex pipelines, and the difference in performance between training and testing periods on original data indicated overfitting (Table 2). The performance scores for the Random Forest method also indicated overfitting (Table 2), even though the complexity of the model was restrained by not making the trees too deep. This also means that the feature importance of the Random Forest method should be interpreted with caution, as feature importance was analysed using the training data and may be less representative for the unseen data. Overall, the non-linear methods Random Forest and TPOT were not superior to the linear multivariate methods LASSO and VAR (Table 2, Fig. 1). A potential explanation to the limited performance of Random Forest and TPOT and their tendency to overfit in this application, is that these methods may require more data to perform better. In addition, these methods are difficult to interpret, limiting their usefulness in practice. The LASSO model showed better performance using the Log_{10}-transformed data (Table 2), thus, this method benefited from reducing skewness in the distribution of the microbial data.

The comparison of the modelled and observed *E. coli* concentrations for the testing period showed that the models managed to capture the seasonal patterns (Figs. 2 and S5), i.e., higher concentrations and more variation during the cold period of the year, but underestimated the magnitude of the peaks. The models that were better at predicting the

Table 2
Performance metrics (mean absolute error – MAE, symmetrical mean absolute percentage error – SMAPE, root mean squared error – RMSE, coefficient of determination – R^2) for training and testing periods for the applied models (Naive, Exponential Smoothing, Autoregressive Integrated Moving Average – ARIMA, Vector Autoregression – VAR, Least Absolute Shrinkage and Selection Operator – LASSO Regression, Random Forest, and Tree-based pipeline optimisation tool – TPOT).

Model	MAE[a]		SMAPE (%)[b]		RMSE[a]		R^2 (–)	
	Train	Test	Train	Test	Train	Test	Train	Test
Original data								
Naive	90	93	66	64	160	159	0.11	0.18
ExpSmoothing	85	85	62	59	143	144	0.28	0.33
ARIMA	84	89	65	66	136	142	0.35	0.35
VAR	66	72	59	64	112	117	0.55	0.56
LASSO	61	69	49	51	110	123	0.57	0.51
Random Forest	36	71	34	52	67	130	0.84	0.46
TPOT	35	65	31	49	63	111	0.86	0.60
Log_{10}-transformed data								
Naive	0.33	0.32	20	17	0.45	0.43	0.15	0.20
ExpSmoothing	0.29	0.28	17	15	0.39	0.36	0.38	0.44
ARIMA	0.28	0.28	16	15	0.38	0.35	0.41	0.45
VAR	0.24	0.24	14	14	0.33	0.32	0.54	0.53
LASSO	0.22	0.22	13	12	0.29	0.29	0.63	0.64
Random Forest	0.17	0.24	10	13	0.23	0.30	0.78	0.59
TPOT	0.22	0.22	13	12	0.29	0.29	0.64	0.62

[a] The unit is (# per 100 mL) for original data and (Log_{10}# per 100 mL) for Log_{10}-transformed data.
[b] Can vary in the range 0–200%.

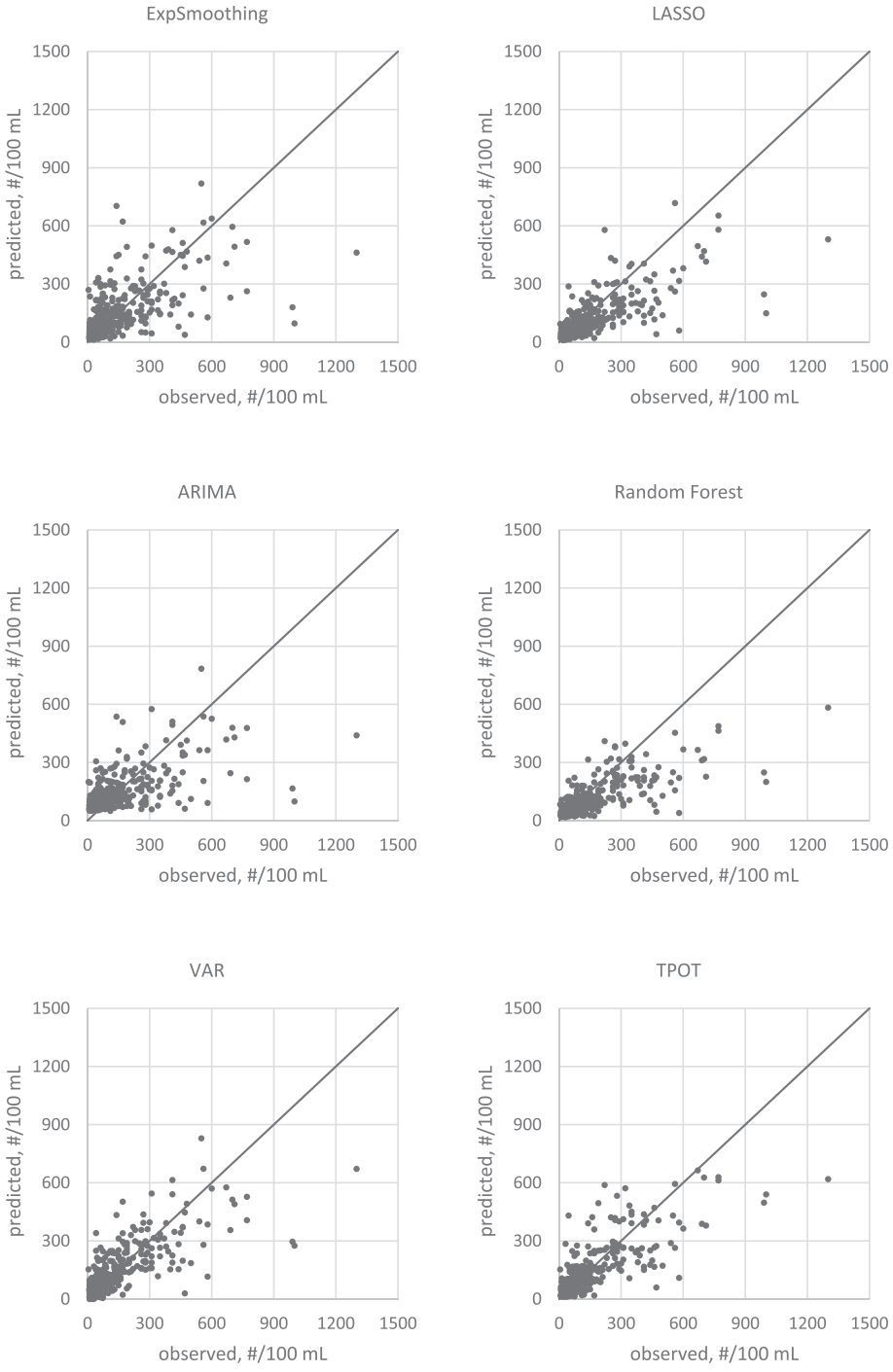


Fig. 1. Scatter plots of observed data ("observed") and modelling results ("predicted") using original data for the testing period for the models Exponential Smoothing (ExpSmoothing), Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Autoregressive Integrated Moving Average (ARIMA), Random Forest, Vector Autoregression (VAR), and Tree-based Pipeline Optimisation tool (TPOT).

peak concentrations, e.g. VAR and TPOT, also had more overestimates (Fig. 1).

The plots of the autocorrelated residuals (Fig. S6) indicated that for some of the models there may be information left in the data that could potentially improve the models. Hyperparameters for the models were selected using cross-validation and grid search, but other parameters could potentially be found resulting in less autocorrelation in the residuals. The residuals for the Random Forest model showed stronger autocorrelation for lower lags and a periodic pattern for higher lags possibly caused by overfitting. The residuals for the LASSO Regression, TPOT and ARIMA models showed no significant autocorrelations, but a slight periodic behaviour for higher lags could be observed. The residuals for the Naive Baseline, Exponential Smoothing and VAR models resembled white noise, but had higher autocorrelation for the residuals at lag 1. Most of the histograms of the residuals (Fig. S6) showed a longer right tail indicating skewness in the distribution of the residuals and the fact that these models tend to underestimate the highest peak values. The histograms of the residuals for the training and testing periods were different for the two models (Random Forest and TPOT) that were prone to overfitting.

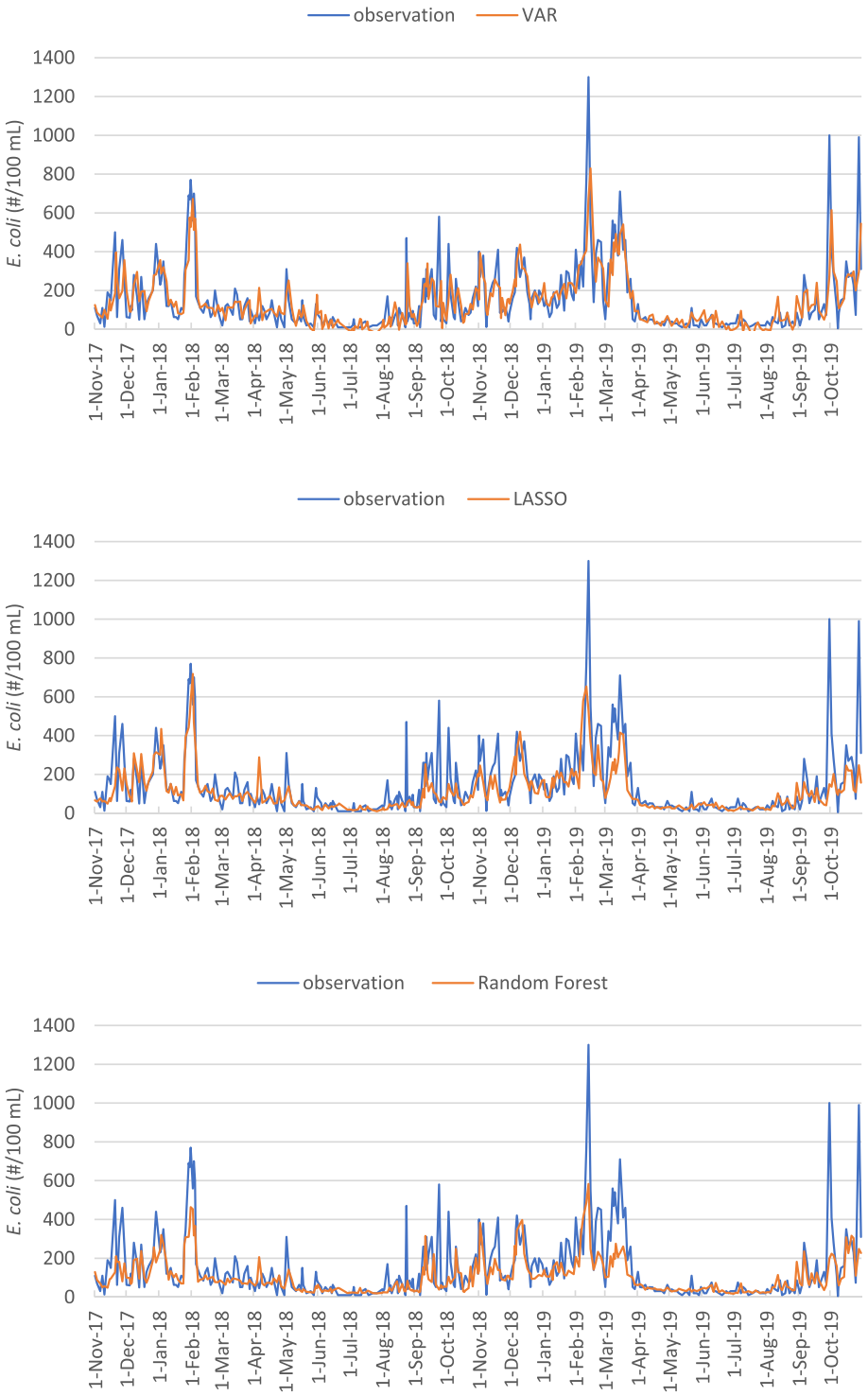


Fig. 2. Time-series of observed data and modelling results using original data for the testing period for the models Vector Autoregression (VAR), Least Absolute Shrinkage and Selection Operator (LASSO) Regression, and Random Forest.

3.2. Model interpretation

Importance of different predictors was analysed using the LASSO Regression and Random Forest models, however, the results for Random Forest should be interpreted with caution. In the case of Random Forest, feature importance is related to how much each feature contributed to the created model. In our case, the Random Forest model had a tendency to overfit, and thus feature importance may not reflect the most suitable features for generalisation on unseen data. The feature importance results for the Random Forest model are included for transparency rather than to draw conclusions.

According to the LASSO Regression and Random Forest models, water temperature, precipitation, and microbial concentrations were important predictors (Fig. 3). The importance of the water temperature is likely a reflection of the seasonality in the microbial data, as water temperature can be regarded as a surrogate of seasonality (Figs. S2 and S3). The *E. coli* concentrations are lower during the warmer part of the year (Fig. S3), this is expected and is mainly due to less runoff

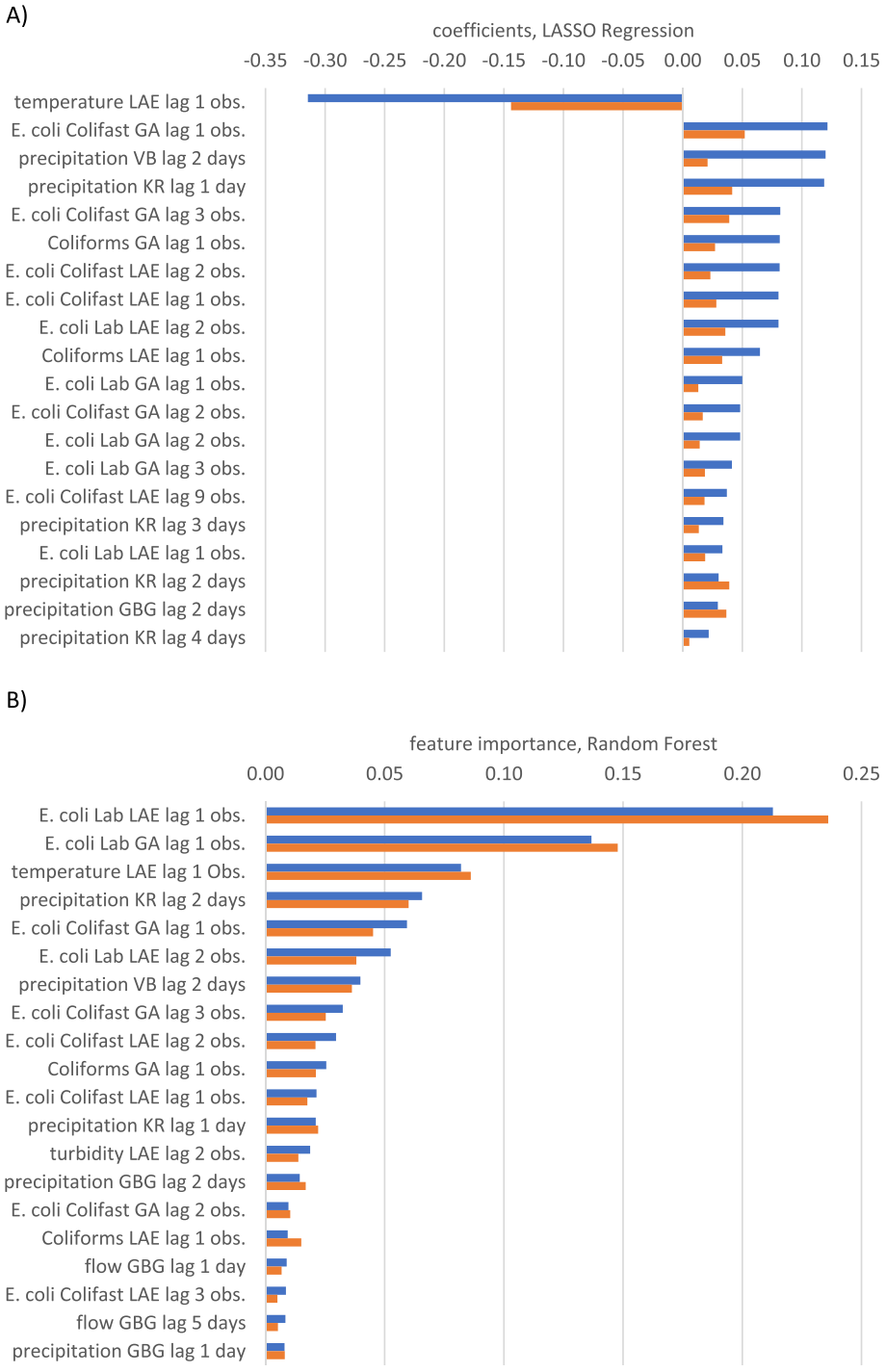


Fig. 3. Coefficients for the LASSO Regression model (A) and feature importance for the Random Forest model (B) using original (blue bars) and Log_{10}-transformed (orange bars) data. The twenty features with the highest absolute values are shown. "Temperature" refers to water temperature; data lags are presented in days or observations ("obs."); the acronyms indicate location (see Fig. S1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and thus less transport of pollution from the sewer system and the diffuse sources in the catchment. A contributing factor is also greater decay of *E. coli* in warmer temperatures and sensitivity to solar radiation. Precipitation at locations upstream was more important than at the water intake; and lags of 1-2 days had higher impact than longer lags, reflecting the relatively short transport times in the river.

E. coli concentrations upstream and at the water intake were important predictors, with the Random Forest model indicating that laboratory data were more important than Colifast data, and the LASSO Regression model indicating the reverse (Fig. 3). The LASSO Regression regularisation path (Fig. S4) showed that when applying a high level of regularisation the predictors used were the *E. coli* laboratory measurements from two different locations. However, in the model that showed the best performance during cross-validation, which had a lower level of regularisation, these predictors did not have the highest effect on the target variable. The magnitude of the coefficients on the *E. coli* laboratory measurement predictors decreased when new predictors were introduced into the model, possibly since these may be highly correlated to already existing predictors. The reason that the Colifast measurements are more important than the more accurate laboratory data

is that the Colifast measurements are more frequent, as the latest Colifast observation represents the afternoon of the previous day, while the latest laboratory observation may represent the situation several days ago. Thus, the Colifast analysis, which is a five-well MPN-method with its relatively high uncertainty of measurement, may improve on predictability compared with the laboratory analysis (using Colilert).

According to the LASSO Regression and Random Forest models, turbidity and water flow were not selected as important predictors in this study. In case of turbidity, its minor importance may be caused by covariation between precipitation and turbidity, as well as relative infrequency of the turbidity measurements (Table 1) included in this study. However, Tornevi et al. (2014) also concluded that precipitation reflects the peaks of faecal contamination better than turbidity, despite having included frequent turbidity measurements (daily mean values). From a practical standpoint, many drinking water treatment plants may have online turbidity data for source water, thus, turbidity should still be regarded an important potential predictor, especially in water sources without cargo shipping or any major causes of turbidity other than precipitation-driven events. River flow is an important factor when making predictions regarding pollutant transport, and the probable reason that flow had only minor importance in this study (Fig. 3) is that geographic location and lags of other predictors indirectly take transport time into account.

3.3. Model applicability and further development

In this project, several models of different complexity and interpretability were compared. The most complex models, i.e. TPOT and Random Forest, appear to slightly overfit the training data, thus, despite the fact that they show high performance on the test data, they cannot be trusted to perform well on future unseen data. If such a modelling approach is to be used in practice, the model interpretability also needs to be considered, as in practice it may be important to understand the contributions of different factors to a prediction. Interpretability is an advantage of simpler models, e.g. LASSO Regression, for which the final model can be setup as a linear equation with the coefficients directly representing the effect of the predictor variables on the target variable.

In further studies, the univariate methods could be improved by addressing the seasonality in the data, for example, using Holt-Winters Exponential Smoothing and Seasonal-ARIMA. However, due to the nature and complexity of the seasonality in the data, other methods may be more suitable, e.g. De Livera et al. (2011). In the multivariate methods, the seasonality was accounted for by including predictors that potentially could explain baseline levels of *E. coli*, e.g. water temperature. Also, the models could be improved by exploring additional imputation methods to address the irregularity in the time-series of microbial data. The issue of the models missing the peak values (Fig. 1) could potentially be addressed by developing a model with the specific purpose of classifying such peaks. In the context of practical application in drinking water production, a model that would also estimate uncertainty of its prediction would be of value, e.g. Bayesian methods (Avila et al., 2018; Panidhapu et al., 2020; Wang et al., 2021). In addition, a longer time horizon for prediction, e.g. several days, would be advantageous from the water management perspective. Furthermore, combining data-driven methods with process-based modelling of water quality (e.g. Dienus et al., 2016; Whitehead et al., 2018) is a promising approach (Young, 1992, 2006).

An optimal forecasting model for managing river water intake should be easily interpreted and make predictions a couple of days ahead based on accessible online data. In the production of safe drinking water, the analysis and evaluation of risks to the entire drinking water system, from the catchment until the water reaches the consumer, is considered of paramount importance by the World Health Organization (WHO, 2017) and the newly revised Drinking Water Directive (EU 2020/2184). During the last twenty years, at-line *E. coli* instrumentation has been key to Gothenburg water intake management. Initially, this was done only at the location of the water intake and thus with limited possibility of forecasting. Expanding with *E. coli* at-line instrumentation upstream the water intake and taking hydrometeorological data into account formed the basis for improved forecasting. Frequent analysis of *E. coli* currently is the main and potentially the only way to discover non-precipitation driven faecal contamination events, while other factors, as precipitation and potentially turbidity, can only predict an increased risk due to wastewater discharges or run-off from grazed agricultural land. Although *E. coli* as a water quality indicator does not represent risk of infection in absolute terms, it does reflect a higher faecal load. Given the overall challenge of predicting microbial concentrations, we may not be able to expect a model to reproduce all peak values. However, from the operational standpoint, discrepancies between the predicted and measured values provide valuable information, as these may indicate events that need intervention, e.g. failures in the wastewater system or diffuse contaminant sources. In this paper, we have demonstrated the potential for improving one day ahead predictions by data-driven models which may be implemented at drinking water treatment plants.

4. Conclusions

- The models that included multiple predictors, i.e. VAR, LASSO Regression, Random Forest, TPOT, performed better than univariate models, i.e. Naive baseline, Exponential Smoothing, ARIMA. External predictors increased the model performance, indicating that some of these models are informative for forecasting *E. coli* concentration at the water intake. The univariate models applied in this study can be further improved, e.g., by addressing seasonality or including more frequent data.
- The most complex models, Random Forest and TPOT, achieved high performance scores on the test data. However, these models showed a tendency of overfitting, indicating that these models would need more data to make better forecasts.
- Among all the models, LASSO Regression and VAR appeared to have the best balance between performance and generalisation. A benefit of LASSO Regression is also feature selection and directly interpretable coefficients showing their effect on the target variable.
- The features that had the largest impact on the target variable *E. coli* concentration at the water intake included: water temperature, microbial concentrations upstream and at the water intake, and precipitation upstream.
- According to the coefficients of the LASSO Regression model, more recent and frequent but less precise data from historical *E. coli* levels play a more important role in the forecast than precise but older data. The feature importance from the Random Forest model indicated a contradictory behaviour, but this can be an effect of overfitting, thus no conclusions can be drawn from the Random Forest model.
- The modelling approach applied in this paper enables the water producers to interpret the measurements in the context of what concentrations can be expected based on the historical data, and thus identify unexplained deviations warranting further investigation of their origin.

CRediT authorship contribution statement

Ekaterina Sokolova: Conceptualization, Visualization, Methodology, Writing – original draft, Project administration, Funding acquisition. **Oscar Ivarsson:** Conceptualization, Visualization, Methodology, Software, Formal analysis, Data curation, Writing – review & editing. **Ann Lilliestrom:** Conceptualization, Methodology, Software, Writing – review & editing. **Nora K. Speicher:** Conceptualization, Methodology, Software, Data curation, Writing – review & editing. **Henrik Rydberg:** Conceptualization, Methodology, Investigation, Writing – review &

editing. **Mia Bondelind:** Conceptualization, Methodology, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was conducted within the research project "ClimAQua – Modelling climate change impacts on microbial risks for a safe and sustainable drinking water system" grant number 2017-01413 funded by Formas – the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning. This research was also funded by CHAIR – Chalmers AI Research Centre. This research was performed within DRICKS – framework programme for drinking water research in Sweden. The water quality data were provided by the City of Gothenburg Department of Sustainable Water and Waste, the water flow data were provided by the power company Vattenfall, and the meteorological data were provided by the Swedish Meteorological and Hydrological Institute.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at https://doi.org/10.1016/j.scitotenv.2021.149798. These data include the Google map of the most important areas described in this article.

References

Akaike, H., 2011. Akaike's information criterion. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 25 https://doi.org/10.1007/978-3-642-04898-2_110.

Asheri Arnon, T., Ezra, S., Fishbain, B., 2019. Water characterization and early contamination detection in highly varying stochastic background water, based on machine learning methodology for processing real-time UV-spectrophotometry. Water Res. 333–342. https://doi.org/10.1016/j.watres.2019.02.027.

Åström, J., Pettersson, T.J.R., Stenström, T.A., 2007. Identification and management of microbial contaminations in a surface drinking water source. J. Water Health 5, 67–79.

Avila, R., Horn, B., Moriarty, E., Hodson, R., Moltchanova, E., 2018. Evaluating statistical model performance in water quality prediction. J. Environ. Manag. 206, 910–919. https://doi.org/10.1016/j.jenvman.2017.11.049.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day.

Bratanova, B., Morrison, G., Fife-Schaw, C., Chenoweth, J., Mangold, M., 2013. Restoring drinking water acceptance following a waterborne disease outbreak: the role of trust, risk perception, and communication. J. Appl. Soc. Psychol. 43, 1761–1770. https://doi.org/10.1111/jasp.12113.

Choi, S.-W., Bae, H.-K., 2018. Daily prediction of total coliform concentrations using artificial neural networks. KSCE J. Civ. Eng. 22, 467–474. https://doi.org/10.1007/s12205-017-0739-y.

De Livera, A.M., Hyndman, R.J., Snyder, R.D., 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. J. Am. Stat. Assoc. 106, 1513–1527. https://doi.org/10.1198/jasa.2011.tm09771.

Dienus, O., Sokolova, E., Nyström, F., Matussek, A., Löfgren, S., Blom, L., Pettersson, T.J.R., Lindgren, P.-E., 2016. Norovirus dynamics in wastewater discharges and in the recipient drinking water source: long-term monitoring and hydrodynamic modeling. Environ. Sci. Technol. 50. https://doi.org/10.1021/acs.est.6b02110.

Francy, D.S., Brady, A.M.G., Cicale, J.R., Dalby, H.D., Stelzer, E.A., 2020. Nowcasting methods for determining microbiological water quality at recreational beaches and drinking-water source waters. J. Microbiol. Methods 175, 105970. https://doi.org/10.1016/j.mimet.2020.105970.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54, 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y.

Laureano-Rosario, A.E., Duncan, A.P., Symonds, E.M., Savic, D.A., Muller-Karger, F.E., 2019. Predicting culturable enterococci exceedances at Escambron Beach, San Juan, Puerto Rico using satellite remote sensing and artificial neural networks. J. Water Health 17, 137–148. https://doi.org/10.2166/wh.2018.128.

Le, T.T., Fu, W., Moore, J.H., 2020. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36, 250–256. https://doi.org/10.1093/bioinformatics/btz470.

Mas, D.M.L., Ahlfeld, D.P., 2007. Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. Hydrol. Sci. J. 52, 713–731. https://doi.org/10.1623/hysj.52.4.713.

Mohammed, H., Hameed, I.A., Seidu, R., 2017a. Adaptive neuro-fuzzy inference system for predicting norovirus in drinking water supply. 2017 International Conference on Informatics, Health and Technology, ICIHT 2017 https://doi.org/10.1109/ICIHT.2017.7899134.

Mohammed, H., Hameed, I.A., Seidu, R., 2017b. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian process for machine learning (GPML) algorithms for the prediction of norovirus concentration in drinking water supply. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). https://doi.org/10.1007/978-3-662-56121-8_4.

Mohammed, H., Hameed, I.A., Seidu, R., 2017c. Random forest tree for predicting fecal indicator organisms in drinking water supply. Proceedings of 4th International Conference on Behavioral, Economic, and Socio-cultural Computing, BESC 2017, pp. 1–6 https://doi.org/10.1109/BESC.2017.8256398.

Mohammed, H., Hameed, I.A., Seidu, R., 2018. Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. Sci. Total Environ. 628–629, 1178–1190. https://doi.org/10.1016/j.scitotenv.2018.02.140.

Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res. 46, 4508–4520. https://doi.org/10.1016/j.watres.2012.05.023.

Panidhapu, A., Li, Z., Aliashrafi, A., Peleato, N.M., 2020. Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. Water Res. 170, 115349. https://doi.org/10.1016/j.watres.2019.115349.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Samanipour, S., Kaserzon, S., Vijayasarathy, S., Jiang, H., Choi, P., Reid, M.J., Mueller, J.F., Thomas, K.V., 2019. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: a proof of concept. Talanta 195, 426–432. https://doi.org/10.1016/j.talanta.2018.11.039.

Seabold, S., Perktold, J., 2010. statsmodels: econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference (SCIPY 2010).

Smith, T.G., 2017. pmdarima: ARIMA estimators for Python [WWW Document]. http://www.alkaline-ml.com/pmdarima (accessed 3.12.21).

Speight, V.L., Mounce, S.R., Boxall, J.B., 2019. Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. Environ. Sci. Water Res. Technol. 5, 747–755. https://doi.org/10.1039/c8ew00733k.

Stevenson, M., Bravo, C., 2019. Advanced turbidity prediction for operational water supply planning. Decis. Support. Syst. 119, 72–84. https://doi.org/10.1016/j.dss.2019.02.009.

Stidson, R.T., Gray, C.A., Mcphail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. 26, 7–18. https://doi.org/10.1111/j.1747-6593.2011.00258.x.

Thoe, W., Wong, S.H.C., Choi, K.W., Lee, J.H.W., 2012. Daily prediction of marine beach water quality in Hong Kong. J. Hydro-Environ. Res. 6, 164–180. https://doi.org/10.1016/j.jher.2012.05.003.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117. https://doi.org/10.1016/j.watres.2014.09.001.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2015. Sunny with a chance of gastroenteritis: predicting swimmer risk at California beaches. Environ. Sci. Technol. 49, 423–431. https://doi.org/10.1021/es504701j.

Tornevi, A., Bergstedt, O., Forsberg, B., 2014. Precipitation effects on microbial pollution in a river: lag structures and seasonal effect modification. PLoS One 9. https://doi.org/10.1371/journal.pone.0098546.

Vijayashanthar, V., Qiao, J., Zhu, Z., Entwistle, P., Yu, G., 2018. Modeling fecal indicator bacteria in urban waterways using artificial neural networks. J. Environ. Eng. (United States) 144. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001377.

Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., Wang, Y., 2021. Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Sci. Total Environ. 765, 142760. https://doi.org/10.1016/j.scitotenv.2020.142760.

Whitehead, P., Bussi, G., Hossain, M.A., Dolk, M., Das, P., Comber, S., Peters, R., Charles, K.J., Hope, R., Hossain, S., 2018. Restoring water quality in the polluted turag-tongi-balu river system, Dhaka: modelling nutrient and total coliform intervention strategies. Sci. Total Environ. 631–632, 223–232. https://doi.org/10.1016/j.scitotenv.2018.03.038.

WHO, 2017. Guidelines for Drinking-water Quality: Fourth Edition Incorporating the First Addendum. World Health Organization, Geneva.

Young, P.C., 1992. Parallel processes in hydrology and water quality: a unified time-series approach. Water Environ. J. 6, 598–612. https://doi.org/10.1111/j.1747-6593.1992.tb00796.x.

Young, P.C., 2006. The data-based mechanistic approach to the modelling, forecasting and control of environmental systems. Annu. Rev. Control. 30, 169–182. https://doi.org/10.1016/j.arcontrol.2006.05.002.